REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	Technical Report	-

4. TITLE AND SUBTITLE
Social Circles Detection from Ego Network and Profile Information

5a. CONTRACT NUMBER
W911NF-14-1-0254

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Paolo Rosso, Roberto Paredes, Jesús Alonso

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Universitat Politècnica De València
Technology Transfer Office_CTT
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
-

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
65349-MA.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This report presents a study making our first approach to the analysis of Social Communities. The experiments conducted provide us with necessary knowledge to produce research within the problem of Social Copying Community Detection, where we will not only consider structural network information
but also the contents of social interactions, with the aim to detect copying communities.

15. SUBJECT TERMS
ego network, social copying community

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Paolo Rosso
UU	UU	UU	UU		19b. TELEPHONE NUMBER 963-877-007e

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Social Circles Detection from Ego Network and Profile Information

ABSTRACT

This report presents a study making our first approach to the analysis of Social Communities. The experiments conducted provide us with necessary knowledge to produce research within the problem of Social Copying Community Detection, where we will not only consider structural network information
but also the contents of social interactions, with the aim to detect copying communities.

First Technical Report of the ARO Project W911NF-14-1-0254

Social Circles Detection
from Ego Network and Profile Information

Paolo Rosso
PRHLT Research Center, Universitat Politècnica de València
https://www.prhlt.upv.es/

19th December 2014

1 Introduction

This report presents a study making our first approach to the analysis of Social Communities. The experiments conducted provide us with necessary knowledge to produce research within the problem of Social Copying Community Detection, where we will not only consider *structural network information* but also the *contents* of social interactions, with the aim to detect copying communities.

In this context, we participated in a competition hosted by Kaggle [4], in which the Social Circles of several ego networks had to be detected based on structural network information and profile features of the users in the net. We tried different machine learning and network analysis techniques to predict the best Circles distribution.

Social Circles have become a useful way of organizing contacts in personal networks. They are therefore currently implemented in the major social networking systems, such as Facebook lists or Google+ circles. Apart from that, they define communities within the users contacts. At this moment, users must define manually their Social Circles. However, research is being conducted on automatic detection of these communities. The latest studies upon this task consider different types of information, combining ego networks structure with information of user profiles [6, 7].

In this report the concrete undertaken task is described. In the next sections we describe the techniques we employed and we discuss the obtained

results. We also provide some brief comments on the following steps of the Social Copying Community Detection project, on the basis of the insights obtained participating in the Social Circles task.

2 Task description

An egonet or ego network consists of the subgraph of the contacts of a particular user (the ego) within the graph of a Social Network. Thus, it includes all the contacts of an ego and the contact relationship between every pair of them. In our study these graphs are undirected. We can further on define Social Circles themselves as subgraphs of the egonet. Egonets may represent overlapping communities (node sharing) or the inclusion of some communities into others (all the nodes of a circle belonging to another). These special characteristics make classical clustering methods inadequate for the task.

The corpus is composed of 110 egonets retrieved from Facebook, with their Social Circles manually labelled by the egos. They are split into two sets, named Training and Test sets, which contain 60 and 50 nets, respectively. All the egonets contain 27,520 users altogether and, for each of these users, information of up to 57 profile features is attached. The ground truth is provided only for Training egonets.

The evaluation measure used at the Kaggle competition, for every egonet, is an edit distance between the ground truth and the prediction with 4 basic operations, every one of them at cost 1:

- Adding a user to an existing circle.
- Creating a circle with one user.
- Removing a user from a circle.
- Deleting a circle with one user.

The evaluation measure of the whole set of predictions is calculated as the sum of the edit distances obtained for all nets.

The competition system allowed for two Test submissions every day, with immediate results over one third of Test nets (public Test). At the competition deadline, on 29th October 2014, the results on the whole Test set (private Test) were released.

3 Methodology

3.1 Prediction techniques

All the techniques that we used for prediction are unsupervised. Therefore, the Training set was only employed to tune the methods.

3.1.1 Multi-Assignment Clustering

Multi-Assignment Clustering (MAC) [11] is a clustering method for vectorial data allowing for the assignment of the different objects into more than one cluster. It provides a probabilistic representation of the decomposition of the data matrix $\mathbf{x}$ into a matrix containing the clusters prototypes $\mathbf{z}$ and a matrix representing the degree to which a particular data vector belongs to the different clusters $\mathbf{u}$:

$$\mathbf{x} = \mathbf{z} \otimes \mathbf{u}, \text{ where } x_{ij} = \bigvee_k [z_{ik} \wedge u_{kj}]$$

In addition, the method models the difference between the original data and the reconstruction made from $\mathbf{z}$ and $\mathbf{u}$ as noise.

The number of circles must be specified, as well as a parameter containing the minimum and maximum number of circles to which a given user can belong. The method is designed to operate with boolean data, although we made some tests with scaled data as well.

3.1.2 Restricted Boltzmann Machines

Restricted Boltzmann Machines (RBM) [10, 1, 3, 2] are stochastic neural networks composed of a visible layer and a hidden layer.

Our experiments were conducted connecting 2 RBMs, being the dimension of the data vectors, d, the number or neurons of the visible layer of the first and the number of circles to predict, c, the number of neurons of the hidden layer of the second. The dimension of the hidden layer of the first RBM is calculated as $h = \sqrt{d * c}$, and it is the same as the dimension of the visible layer of the second. The activation function was defined as sigmoid for the first RBM and as either linear or sigmoid for the second RBM.

This system was trained with our data and after that, the data vectors were passed through the 2 RBMs system. In the case of linear activation at the second RBM, a threshold must be defined and the outputs surpassing that threshold are interpreted as memberships of the correspondent circle. In the case of sigmoid activation, only exact 1s are interpreted as circle memberships.

3.1.3 k-Clique Communities

This technique does not make use of the profile features information, considering only the contact relationships between users. A k-clique is a complete subgraph of size k. A k-clique community is defined as a union of all k-cliques that can be reached from each other through a series of adjacent k-cliques [9].

The percolation algorithm used to infer k-clique communities is exponential, which makes this technique unfeasible when treating egonets with a large number of users. A possible solution is to define a time limit for the detection of k-cliques. Otherwise, this particular egonets may be modelled by another method or left with a Singleton Circle submission.

We tested the performance of several values of k. The values around $k = 7$ produced the best results and the Social Circles obtained with them were the ones ranking highest in the competition.

3.2 Data Manipulation

The tested techniques need a matrix as input and so diverse matricial representations of both contact relationship and profile features information were designed. In all representations, the matrix rows represent the users in an egonet.

For the contact relationship information the columns represent users as well, and a value distinct from 0 indicates a relationship between the row user and the column user. We defined three relationship ranks:

- 1st rank: Friends.
- 2nd rank: Friends of friends.
- 3rd rank: Friends of friends of friends.

Some representations contain information of several ranks. In that case, they may remain in different columns or be aggregated into only one. In addition, the information may be binary or scaled depending on the friendship rank. For example, one representation contains information of the three levels of friendship, giving a value of 1 to the 1st rank relationship, 0.5 to the 2nd rank relationship and 0.25 to the 3rd rank relationship. Later, the three values are aggregated into one column resulting the maximum of them.

Concerning the user profile features, out of the 57 appearing in the corpus, in the end we used only 3. The reasons are that some of them are very seldom informed whereas others are redundant, like the hometown ID and

the hometown name, and others are not considered to be relevant for the task, such as the first name. The features we used are: hometown ID, schools IDs and employers IDs.

As had been done with structural information, several matricial representations of the features have been created. Columns may represent feature values or users. In the first case, the representation shows whether the set of values of the row user for a particular feature includes the column value. In the second case, it shows whether the row user and the column user share a value of a particular feature. This can be further aggregated into only one column by scaling the values depending on the number of shared features. For instance, 1 if they share the 3 features, $\frac{2}{3}$ if they share 2 features, and $\frac{1}{3}$ if they only share 1 feature.

3.3 The Singleton Circle Benchmark

Due to the evaluation measure, the penalty of misclassifying users into circles is greater than that resulting from no classification. For this reason, the submission of no circle should give a greater performance than benchmarks like Connected Components and All Friends in One Circle. As the evaluation system requires at least a non-empty circle, we defined a benchmark using an only circle with one random member. This was further refined by choosing the member with the highest Page Rank value [8, 5].

Some techniques may not have a good performance predicting Social Circles for certain egonets, while improving the results for the rest. The worst outputs may then be substituted with the Singleton Circle in order to obtain better results.

4 Results

The best results produced by the described methods on the different datasets are shown in the following table. Smaller values of the evaluation measure mean a more reduced number of basic operations needed to transform the prediction into the ground truth. Therefore, the smaller the value of the evaluation measure, the better the results.

In particular, the results obtained by k-Clique Communities allowed us to reach the 11th position of the Public Leaderboard and the 61st of the Private Leaderboard, over 203 total participants. In the following table we show our classification in the Private Leaderboard, in comparison with the 3 highest and the 3 lowest results:

Method	Eval. measure		
	Train. Set	Public Test	Private Test
Singleton Circle	17101	3004	8625
MAC	17011	4461	10026
RBM	15829	3000	8556
k-Clique Comm.	15350	2633	7999

Position	Team	Eval. measure
1	tom denton	6637
2	Adrien	6665
3	Misha Siverski (PZAD, MSU, Russia)	6710
...		
61	**PRHLT4ARO**	7999
...		
201	Justin123	12921
202	yang	13305
203	wjwolf	25341

Although our approach did not rank among the top positions, the difference between our result and the ones produced by the best ranked teams is considerably smaller, in comparison, than the difference between our result and the ones produced by the lowest ranked teams.

5 Discussion

In this section we summarize the main insights we gained participating in Kaggle's task on Social Circles Detection.

(i) The data representations employed are sparse. This may be problematic when considering RBMs. This inconvenient was positively solved implementing a sparsity treatment with the RBM algorithm.

(ii) The ground truth was labelled by the egos themselves. We suspect there may be cases of incomplete or inaccurate labelling. In addition, most profile features are highly sparsely informed, which makes their use in our techniques difficult. These seem to be the reasons why network analysis or graph theory techniques have performed better than machine learning algorithms.

(iii) Of all the techniques, MAC has obtained the worst performance, beating the Singleton Circle Benchmark only for the Training Set although by a lower margin. Apart from the reasons described in (ii), it might not be an adequate method for this task. RBMs have performed better, beating the Singleton Circle in every case, although the best results were the ones produced by k-Clique Communities, on every dataset.

6 Further work

6.1 Social Copying Community Detection

Due to our participation in the Kaggle competition, we have become acquainted with the problem of Community Detection in graphs. The knowledge acquired in the course of the competition is highly valuable and will be a helpful aid in our research in the topic of Social Copying Community Detection. Nevertheless, unlike Social Circles Detection, based mostly on *structural network information*, Social Copying Community Detection will rely on the *content* of user publications as a source of information, as well.

Our study will be performed on Twitter, with the following methodology:

(i) We download the dataset (streaming tweets).

(ii) We identify clusters within the data graph whose nodes have a high content similarity. A crowdsourcing will provide us with a gold standard for the validation of this segmentation.

(iii) We apply the previously described algorithms to detect copying within these communities.

We believe that the results provided by RBMs should have ranked higher on the Social Circles Detection leaderboard in an alternative scenario with correctly labelled circles and comprehensive profile features, as mentioned in the previous section. We will, therefore, investigate RBMs further for Social Copying Community Detection and we feel confident they will be an appropriate algorithm to extract good results from structural as well as content information.

6.2 Crowdsourcing

We are currently downloading our dataset, composed of a stream of tweets focused on particular locations. We are especially interested in detecting the

major influencers and streaming their followers' tweets, as we believe this should increase the appearance of copies. Once the data is collected, the annotators will provide a similarity measure between certain pairs of tweets.

References

[1] Yoav Freund and David Haussler. *Unsupervised learning of distributions of binary vectors using two layer networks*. Computer Research Laboratory [University of California, Santa Cruz], 1994.

[2] Geoffrey Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010.

[3] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002. MIT Press.

[4] Kaggle. Learning social circles in networks. `http://www.kaggle.com/c/learning-social-circles`.

[5] Amy N Langville and Carl D Meyer. A survey of eigenvector methods for web information retrieval. *SIAM review*, 47(1):135–161, 2005. SIAM.

[6] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[7] Julian McAuley and Jure Leskovec. Discovering social circles in ego networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(1):4, 2014. ACM.

[8] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999. Stanford InfoLab.

[9] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005. Nature Publishing Group.

[10] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986. Department of Computer Science, University of Colorado, Boulder.

[11] Andreas P Streich, Mario Frank, David Basin, and Joachim M Buhmann. Multi-assignment clustering for boolean data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 969–976. ACM, 2009.